Discovering the genetic architecture of the mind

Karlsson Linnér, R.

2019

document version
Publisher's PDF, also known as Version of record

Discussion and conclusion

"Life should not be a journey to the grave with the intention of arriving safely in a pretty and well preserved body, but rather to skid in broadside in a cloud of smoke, thoroughly used up, totally worn out, and loudly proclaiming "Wow! What a Ride!"
Hunter S. Thompson

"My supervisor once told me that the only good thesis is a finished thesis."
Philipp D. Koellinger

6

In this sixth and final chapter, I first discuss how selected results answer the main research questions and relate to the previous literature. Thereafter, I conclude the thesis and discuss some study limitations. Finally, the thesis is abbreviated in a general summary.

1. a. Do GWAS in hundreds of thousands of individuals identify robustly associated SNPs with the main phenotypes studied in Chapters 2–3?

The genome-wide association analyses identified a varying number of SNPs associated with the main phenotypes studied in **Chapters 2–3**. It was expected that the number of identified associations would depend strongly on sample size. The analyses of subjective well-being, depressive symptoms, and neuroticism are the smallest of the main GWAS, performed in 298,420, 161,460, and 170,911 individuals, respectively. Upon publication, while these were the largest GWAS reported for these traits[1] the sample sizes were only sufficient to identify a handful genome-wide significant associations. We found three approximately independent "lead SNPs" associated with subjective well-being, two with depressive symptoms, and 11 with neuroticism (**Table 3.1**).

It is possible that the association signal was attenuated by suboptimal phenotypic measurement and overlap, as well as imperfect genetic correlations, across the 59 study cohorts included in the meta-analysis of subjective well-being. In contrary, the meta-analyses of depressive symptoms and neuroticism included only two to three study cohorts, which may explain the similar number of hits, while those GWAS were much smaller. We estimated the SNP heritability of neuroticism to be about twice that of subjective well-being and depressive symptom, 9.1% versus 4–4.7%, which likely explains why the number of associations is the largest for neuroticism. Nonetheless, the identified lead SNPs are among the first genome-wide significant associations reported for these traits[2,3], and the findings could be considered a small first step in the effort to identify the genetic variants responsible for their missing heritability.

In **Chapter 2**, the GWAS identified a much larger number of lead SNPs; 124 with general risk tolerance, 167 with adventurousness, 42 with automobile speeding propensity, 85 with drinks per week, 223 with ever smoker, 117 with number of sexual partners, and 106 with the first PC of the four risky behaviors (**Supplementary Materials**). A few previously reported associations are corroborated by the findings, while the vast majority of the identified associations are novel. The analysis of general risk tolerance is the largest of the GWAS, which included 939,908 individuals, and the six supplementary GWAS included 315,894 to 557,923 individuals. Compared to **Chapter 3**, larger sample size is an obvious explanation of the much greater number of identified lead SNPs. However, two additional explanations deserve to be mentioned.

First, most of the GWAS in **Chapter 2** were either performed in a single large and relatively homogenous cohort, the UKB, or as a meta-analysis of two large and relatively homogenous cohorts (the UKB and the 23andMe cohort). Therefore, it is likely that there was less attenuation because of phenotypic heterogeneity and imperfect genetic correlation. Second, the GWAS in **Chapter 2** were performed with a recently developed method and software with which it is possible to perform computationally tractable GWAS with linear mixed models in large genetic datasets[4,5]. Importantly, GWAS performed with linear mixed models have increased power to detect genetic associations. This method was introduced in the later stages of the study reported in **Chapter 3**, and was therefore not implemented there. In combination, these factors may explain why the GWAS in **Chapter 2** identified a much larger number of genetic associations.

We performed various types of replication and robustness analyses, which suggest that the results are not spurious overall. Thus, our findings show that GWAS in hundreds of thousands of participants can identify robust associations with mental and behavioral traits for which it has proven difficult to identify genes and biomarkers[6,7]. Overall, our findings motivate further

genetic studies of these traits when larger samples become available in order to discover the remaining genetic variants responsible for their missing heritability, discussed next.

b. What do the results reveal about their genetic architectures?

In **Chapter 2**, the lead SNPs account for only a small part of the missing heritability. In sum, the 124 general-risk-tolerance lead SNPs explain about 0.5% of the phenotypic variation. In that chapter, we estimated the SNP heritability of general risk tolerance with high precision to between 5.5 and 8.5%, depending on the method. Thus, the size of the hiding heritability, which is expected to be revealed by increased GWAS sample size, is in the order of 5–8%. The same applies for the supplementary phenotypes. The joint R^2 of the lead SNPs, depending on the trait, is about 0.41 to 1.91%, and the hiding heritability is between 5.8 and 15.9%. In **Chapter 3**, the handful of lead SNPs jointly account for an even smaller, almost minute, part of the variability, roughly 0.03% of the variation in subjective well-being, 0.06% in depressive symptoms, and 0.23% in neuroticism. Thus, the hiding heritability remains almost equal to the SNP heritability.

Polygenic prediction performed with a much larger set of SNPs (about 1 million) supports the notion that the hiding heritability will be revealed with increased sample size. A general-risk-tolerance PGS explains up to 1.8% of the phenotypic variation, about 3.6 times more than the 124 lead SNPs. A PGS for subjective well-being explains about 0.9% of the variation in subjective well-being, 0.5% in depressive symptoms, and 0.7% in neuroticism. Thus, the PGS for subjective well-being explains about 30 times more than the three lead SNPs. Unfortunately, GWAS in hundreds of thousands to almost a million individuals did not reveal the bulk of the missing heritability. Nonetheless, the following paragraphs will discuss some interesting features of the trait's genetic architectures that were revealed by the results.

As discussed in **Chapter 1**, it has been proposed that part of the still-missing heritability[8], which is not expected to be revealed by increased GWAS sample size, may be explained by structural variants[9,10], such as inversions. Overall, it has not been concluded whether the alleles of structural variants are actually in LD with SNPs, and thus, tagged in GWAS[11]. Interestingly, we found that neuroticism is strongly associated ($P < 2.01\times10^{-15}$) with a previously known inversion located on chromosome 8 (~7.89 to 11.8 Mb), and less strongly ($P < 1.25\times10^{-9}$) with another known inversion on chromosome 17 (~43.6 to 44.3 Mb). The estimated effect of the chromosome 8 inversion is stronger than that of any individual lead SNP ($R^2 \sim 0.06\%$), but similar to the joint R^2 of the six near-independent and significant SNPs that tagged it. The effect of the chromosome 17 inversion was estimated to 0.033%, and thus, is similar to the effect of its only significant tag SNP ($R^2 \sim 0.028\%$). Both inversions were estimated to be bi-allelic and common in the British population, with minor allele frequencies of about 0.44 and 0.22, respectively. It is obviously too early to say for sure, but because the effects of these inversions appear to be tagged by common SNPs, and would thus be identified by traditional GWAS, it remains an open question whether such *common* structural variants actually account for the still-hiding heritability, instead of the missing heritability. At the same time, rare structural variants that cannot be tagged in traditional GWAS because of limited linkage with common or low-frequency SNPs[12,13] would instead contribute to the still-hiding heritability.

Notably, we annotated many of the lead SNPs identified in **Chapter 2** to more than a hundred smaller structural variants of various types identified by Sudmant et al. (2015), as well as to about 40 larger candidate inversions identified by Gonzalez & Esko (2017, unpublished), including the two discussed above. Because of the very large number it was not within the scope of that study to perform the same thorough variant calling as in **Chapter 3**. Nonetheless, the annotated variants could be considered interesting candidates for future studies on the relation between mental traits and structural variation. A surprising finding was that many of these regions harbor great overlap of association signal across the main phenotypes. In the literature,

there is conflicting evidence with respect to whether structural variants have much influence on complex traits[11,14]. Our results, and in particular the more thorough variant calling in **Chapter 3**, support the notion that common SNPs do indeed tag structural variation, and the findings contribute to the somewhat limited evidence that suggests that common structural variants indeed have an influence on complex traits and disorders.

Next, the findings align with the previous literature that has found that most GWAS findings are located in non-coding intergenic regions[10,15,16]. In **Chapter 2**, about 64% of the identified lead SNPs are intergenic[17]. These findings suggest that many of the identified SNPs are likely to be involved in the regulation of gene expression rather than the structure of RNA and/or proteins[16]. Because of the large number of identified associations in **Chapter 2**, it was not within the scope of that study to closely examine the genomic function of each association. Next, the 16 lead SNPs identified in **Chapter 3** tell a slightly different story because almost all are intragenic[17], and thus, located within genes. 14 of those are intron variants, i.e., within genes and involved in RNA transcription, but that get spliced away before translation into protein[18]. The remaining two are intergenic, and none are coding variants. However, a few of the 16 lead SNPs appear to be in linkage with nonsynonymous coding variants, for example the neuroticism lead SNP that tagged the inversion on chromosome 17. That particular SNP is in strong linkage with 11 known missense variants. It is possible that those are the truly causal variants in that region, but more extensive fine-mapping is required to come to a firm conclusion. Overall, fine-mapping of the associations identified in **Chapters 2–3** is an interesting avenue for future research[19].

Finally, the small effects that we identified for most of the main phenotypes, and the difference in R^2 between the lead SNPs and the PGS constructed with about a million variants, align with the previous evidence that suggests that mental traits are highly polygenic and typically lack common variants with large effects. An exception is the large effect of the top association with drinks per week, located in the *ADH1B* gene. It is reassuring that we identified several associations between drinks per week and various members of the *ADH* gene family because of their well-known function in alcohol metabolism[20,21]. Yet, it remains to be decided whether these traits confer to a completely infinitesimal "omnigenic" model[22]. Ultimately, it could be that we may never be able to detect all truly causal SNPs because an infinite sample size would be required[23]. Identification of the absolutely smallest effects could still be relevant to fully understand trait biology. Yet, I think that the absolutely smallest effects would not really contribute much to the predictive power of a polygenic score because they would not be estimated with high-enough precision in any finite sample. However, an observation in **Chapter 2** speaks against the premise that all common SNPs are causal. Namely, the optimal predictive power with LDpred[24], a polygenic-score method that includes SNPs in linkage disequilibrium (LD) and adjusts their effects for non-independence, was achieved when the fraction of causal SNPs was assumed to be 30%. However, that evidence is weak due to several limitations, such as the non-asymptotic sample size of the discovery GWAS and possible inaccuracy in the reference panel LD. Overall, it is too early to conclude whether the genetic effects on the studied traits are completely infinitesimal.

c. What is the potential to use the results to strengthen inference in empirical research and to perform multi-trait analyses of genetically correlated traits?

In **Chapter 2**, the potential to strengthen inference in empirical research was evaluated through polygenic prediction of alternative measures of risk tolerance, as well as personality dimensions and various risky real-world behaviors, among other traits. We found that a general-risk-tolerance PGS significantly explains variation in many of these traits. However, the predictive power was limited. It is the greatest for general risk tolerance itself, ranging from about 1 to 1.8%, and typically less than 1% for the other traits. Overall, the predictive power could be

considered disappointing, for example in comparison with a recent large-scale GWAS of educational attainment where the predictive power reached 13%[25], though it should be noted that the SNP heritability of general risk tolerance is less than that of educational attainment. Nonetheless, it is possible that a PGS with the current predictive power could already be beneficial to strengthen inference in empirical research of risk-related outcomes, discussed next.

Rietveld et al. (2013) proposed a framework to evaluate the contribution of a PGS to empirical research[26]. The basic idea is that a PGS can increase the power of a randomized control trial by decreasing residual variation. Thus, the inclusion of a PGS could offset a reduction in experimental sample size, which can be translated into financial terms. A financial break-even point can be approximated as $(N_a - N_b) \times P = N_b \times G$, where N_a is the originally number of study participants, N_b is the number of study participants in the reduced sample, P is the original cost per study participant, and G is the cost to genotype each of the study participants in the reduced sample. Thus, the left-hand side of the equation represents the cost saved, while the right-hand side represents the additional cost to genotype the remaining study participants. Rietveld et al. (2013) give examples of expensive educational reforms that can amount to more than $10,000 per participant and year to evaluate. In such cases, it is possible that even a PGS with modest predictive accuracy can substantially reduce the study cost.

But at what P could our general-risk-tolerance PGS, which explains 1.8% of the outcome variation, be motivated? To illustrate, assume that a prospective randomized control trial intends to include 300 participants (i.e., N_a) to detect a treatment effect of 5% (R^2) on some risk-related outcome. In that case, the theoretical study power is 97.5%. Under these assumptions, calculations show that power can be kept constant with about 295 participants (i.e., N_b), that is five participants less. Today, it could be reasonable to assume that the cost of genotyping (G) is €50 per study participant[x]. In this case, P is equal to €2,950. Thus, our general-risk-tolerance PGS could already be beneficial to some expensive randomized control trials. However, there are also many conceivable studies that are not expensive enough to motivate the inclusion of a PGS based only on financial motives. In such cases, our PGS could be used anyways to increase power and/or control for unobserved heterogeneity. This particular example assumed the inclusion of a single PGS, but once the participants have been genotyped it is easy to include multiple polygenic scores at a low marginal cost. In summary, I argue that our results have the potential to strengthen inference in empirical research, and in addition, can be financially beneficial to particularly expensive experiments of risk taking.

In **Chapter 2**, genetic correlations were estimated between general risk tolerance and about 30 other traits. In summary, strong correlations were identified with many risky lifestyle behaviors, and weak to moderate correlations were identified with several socioeconomic and neuropsychiatric traits, and with personality dimensions. In that chapter, multi-trait analyses with the summary statistics already suggest a benefit of multi-trait analysis. Using MTAG, we leveraged information from the supplementary phenotypes and were able to increase the number of general-risk-tolerance leads SNPs to 312, which jointly explain 0.93% of the phenotypic variation. Thus, almost twice that of the 124 lead SNPs identified in the discovery GWAS. It was also the general-risk-tolerance PGS based on the MTAG results that had the greatest predictive power. Future studies should be to able to use our results to boost power in multi-trait analyses of genetically correlated traits. Such decisions can be guided by the genetic correlations we estimated. Similarly, we estimated strong genetic correlations between subjective well-being, depressive symptoms, and neuroticism with anxiety disorders. There are currently few reported genome-wide associations with anxiety disorders[27], and thus, it could be

[x] Gencove. Retrieved September 13, 2018, from https://gencove.com/

beneficial to use our results to aid the identification of genetic associations with that condition. Finally, proxy-phenotype analyses in **Chapters 2–3** identified plausible candidate associations with several traits, such as ADHD, lifetime cannabis use, and self-employment, in loci that have not been previously reported for these traits. In conclusion, our results can be leveraged for multi-trait analysis of genetically correlated traits.

c. What do the results reveal about biological mechanisms and pathways?

We performed bioinformatic analyses of general risk tolerance, subjective well-being, depressive symptoms, and neuroticism with the GWAS summary statistics. Importantly, functional partitioning with stratified LD Score regression found the strongest enrichment for all four traits in the functional category central nervous system. That is encouraging because these phenotypes are foremost expected to be related to brain function. It would be worrying if the strongest signal would have been found in seemingly unrelated pathways and tissues. Also, a weaker, but yet significant level of enrichment was found for general risk tolerance in the category immune/hematopoietic (while excluding the MHC region). To the best of our knowledge, this is one of the first times a study finds significant enrichment for a mental trait in the category immune/hematopoietic, though this has previously been implicated for some mental disorders. Next, the functional category adrenal/pancreas was found enriched for subjective well-being and depressive symptoms, which aligns with previously reported associations between depression and the stress hormone cortisol[28], which is produced in the adrenal cortex. Interestingly, the lack of enrichment for general risk tolerance in the category adrenal/pancreas contests previously reported associations with cortisol[29].

Next, Gene Network analysis strongly suggest the involvement of the following brain regions in general risk tolerance: prefrontal cortex, frontal lobe, visual cortex, parietal lobe, and putamen. The biological pathways that are implicated by that analysis include particular brain-related mechanisms, such as dendritic and synaptic processes. That method, together with DEPICT, mutually implicate the involvement of both glutamate and GABA neurotransmitters. However, neither SMR nor competitive gene-set analysis support that finding. For brevity, the interested reader can find a detailed discussion of this conflicting result in the **Supplementary Materials**. Notably, few to no other large-scale GWAS of mental traits have reported evidence in favor of the involvement of *both* glutamate and GABA.

In **Chapter 3**, Gene Network analysis of subjective well-being suggest the involvement of particular brain regions, such as putamen, thalamus, and the visual cortex, but in contrast to risk tolerance, that analysis also ranks many seemingly unrelated tissues among the highest. That result is likely explained by the smaller sample size of the GWAS of subjective well-being, and future studies in larger samples should be able to more accurately pinpoint particular brain regions. Overall, these bioinformatic analyses of GWAS summary statistics are limited by the availability of external reference data, and it may be useful to revisit these analyses in the future as more reference data becomes available.

d. What do the results suggest about the biological pathways and candidate genes that have previously been hypothesized to influence risk taking?

As discussed in **Chapter 1**, previous studies of biological pathways and candidate genes in relation to risk taking could be considered hampered by several methodological shortcomings, such as small sample size. Because this is by far the largest genetic study of risk tolerance it provides an excellent opportunity to replicate the previously hypothesized pathways and genes. Previously, five pathways have frequently been hypothesized to influence risk tolerance: testosterone and estrogen, dopamine and serotonin, and cortisol. Competitive gene-set analysis did not find significant enrichment for neither of the five, which is further supported by the results of Gene Network analysis and DEPICT. Next, we investigated 15 previously

hypothesized candidate genes, and found that none are among the 285 significant genes identified in gene-based analyses with MAGMA. Thus, a large number of other genes appear to be more important than those that have previously been suggested in the literature. Finally, we investigated whether particular SNPs located within, or that are known to tag, those 15 candidate genes are associated with general risk tolerance. Overall, no such SNPs replicate at genome-wide significance. In summary, our results contest the involvement of the biological pathways and candidate genes that have previously been hypothesized to influence risk taking.

2. Is educational attainment associated with CpG methylation and how does the strength of association compare to biologically proximate factors?

In **Chapter 4**, the EWAS, which included 10,767 individuals, identified that educational attainment is associated with 37 CpG probes at epigenome-wide significance. Nine of those probes remain significant in an adjusted model that controls for BMI and smoking. Thus, when we considered two biologically proximate confounders then the number of associations dropped quite drastically. We also observed that there is much less inflation of the overall test statistic in the adjusted model, which is a sign that these covariates are important to avoid inflation, and thus, possible omitted variable bias across most or all probes. We benchmarked the 50 top probes from the adjusted model in comparison to four biologically proximate factors and found that their effect sizes are many times weaker compared to the strongest effects reported for smoking and maternal smoking, a few times weaker compared to alcohol consumption, and about 50% weaker compared to BMI. In addition, most of the 44 probes with association P value less than 1×10^{-5}, in the adjusted model, have previously been associated with smoking and/or maternal smoking[30,31].

Disconcertingly, the results of several robustness analyses suggest that most, if not all, of the identified associations are driven by confounding from smoking even though we controlled for smoking with a covariate. It appears as if that covariate is not detailed enough to fully account for the strong and long-lasting effects from smoking on methylation. Typically, the smoking covariate available in most epidemiological cohorts is coarsely measured, which translates into measurement error, and thus, residual confounding[32]. Unfortunately, alcohol consumption was not available or consistently measured across the included study cohorts. Therefore, we cannot rule out that alcohol consumption could be responsible for the remaining inflation of the overall test statistic; λ_{GC} is about 1.06 after genomic control (adjusted model). Notwithstanding the disappointing findings, our study contributes by being many times larger than most previous studies in behavioral epigenetics, and it can hopefully increase the awareness of the many issues surrounding confounding from lifestyle factors in observational studies of methylation.

What do the results suggest about the hypothesis that psychosocial experiences influence disease via methylation?

Overall, our results do not support the hypothesis that a major life experience would be strongly associated with methylation, or at least in a direct manner that cannot be attributed to biologically proximate factors. Thus, it is questionable if psychosocial experiences actually influence disease via methylation. To my knowledge, there are no studies that have investigated that hypothesis with a randomized design in humans[33,34], which is an important but potentially infeasible avenue for future research. Thus, before causality has been established, a more parsimonious explanation is that the reported associations between psychosocial experiences and methylation are driven by a range of biological confounders. Notably, several review articles on the subject acknowledge that possibility[35–38]. Nevertheless, I find it alarming that many studies in behavioral epigenetics brush aside such concerns in favor of more sensational claims, which may lead to a legitimization of bad research practices, solidify the trust in weak results, and add to the overall hype of epigenetic epidemiology[33]. In my opinion, if the goal of

epigenetic epidemiology is indeed to provide biomarkers and guide health interventions and social policies, which is a frequently voiced motivation[34,39,40], then carefulness must be the norm for the execution and interpretation in behavioral epigenetics. Otherwise, there is a risk that resources are misallocated to ineffective medical and social interventions, non-replicable studies, and that clinical trials may experience similar failure rates as those observed in industry-conducted trials based on published findings from cancer research[41].

On the introductory page of **Chapter 4**, I quote Carl Sagan on the virtue of keeping an open but not too open mind. I chose that particular quote because it relates to my skepticism towards the notion that "biological embedding of psychosocial stress"[34,39,42] would be a strong explanatory factor of health inequalities, rather than positive health behaviors, lifestyle factors, and underlying genetic liability towards multiple of these factors. I speculate that this may be a politically appealing explanation because it implies less blame and shame of the health inequalities across socioeconomic strata[43,44]. However, good intentions do not motivate bad research practices. Importantly, I am not saying that stress, bad working conditions, and poverty are not major negative health factors[6,45,46], I merely question the need to hypothesize about a largely unobservable social stress proxied by higher-order factors such as socioeconomic status, which by design becomes prone to being confounded by unobserved heterogeneity.

3. **What is the overall quality and evidential value of the previous studies that have tested antisocial behavior for association with an interaction effect between adverse life events and *5-HTTLPR*?**

None of the eight reviewed studies attained a completely satisfying score in the quality assessment. Importantly, all the studies fail to account for all relevant covariate×gene and covariate×environment interaction terms[47]. Apparently, the studies disagree on what covariates are important in this setting because the included covariates vary a lot. Thus, omitted variable bias could have inflated the strength of association in some of the studies. Further, none of the studies control for population stratification with molecular genetic data and several of the studies analyze samples with admixed ancestry. Therefore, none of the studies can rule out the alternative explanation that other genetic and environmental factors could be attributable. Thus, the quality of the eight studies could be considered weak. Evidently, the multiple sources of bias that we identified severely limit the evidential value of the subset of studies that report a significant interaction. Also, more recent GWAS findings with respect to antisocial behavior[48], and our GWAS of the genetically correlated trait general risk tolerance ($r_g \sim 0.35$), do not provide support of a main effect of *5-HTTLPR*. The absence of a direct effect is a weaker but yet important indicator that it is unlikely that *5-HTTLPR* would be a moderator of the effect of adverse life events. In conclusion, the findings contest the hypothesis that antisocial behavior would be associated with an interaction between adverse life events and *5-HTTLPR*.

Conclusion

The central purpose of this thesis has been to investigate the genetic architecture of the main phenotypes. **Chapters 2–3** found that GWAS in hundreds of thousands to almost 1 million individuals have the ability to identify robust genetic associations with general risk tolerance, adventurousness, and the four risky behaviors and their first PC, as well as with subjective well-being, depressive symptoms, and neuroticism. Many of the identified associations were the first reported for these traits. However, the identified variants explain only a small part of the hiding heritability. Nonetheless, the results uncovered several interesting aspects of their genetic architectures, such as genetic overlap in a couple of inversion polymorphisms, and the thesis has shown that the results can be used to strengthen inference in empirical research and to perform multi-trait analyses of genetically correlated traits. An interesting finding is the novel

lead about the possible involvement of glutamatergic and GABAergic neurotransmission in relation to general risk tolerance, and the lack of support for the involvement of the previously hypothesized genes and biological pathways. **Chapter 4** found that educational attainment is not strongly associated with methylation, and that several biologically proximate confounders are major concerns for the robustness of observational studies that test whether psychosocial experiences are associated with methylation. Finally, **Chapter 5** found that the overall quality is low for the previous studies that have tested antisocial behavior for association with a gene-environment interaction between adverse life events and *5-HTTLPR*, and that there is little evidential support in favor of that interaction.

Study limitations

The vast majority of the study participants are of European descent. Thus, it is not certain that the results can be generalized to populations of different ancestry. Overall, larger genetic studies in non-Europeans are warranted[49,50]. Next, it was possible to amass very large samples because relatively simple self-reported measures were studied. That, in combination with phenotypic heterogeneity across the study cohorts, could have attenuated the power to detect genetic associations. Although some of the studies used repeated measurements most analyses were performed cross-sectionally. Out of the four studies, I deem that a longitudinal study design could have benefited the EWAS the most. In that study, a longitudinal design could have led to more accurate inference of whether associations with CpG probes would have appeared before or after initiation of education as well as smoking initiation. Also, because methylation is dynamic and unstable, repeated measurements could have been used to correct for measurement error. Unfortunately, we could not identify any large-enough EWAS sample with quasi-random variation that would allow us to investigate causality.

Furthermore, the GWAS reported in this thesis focused on common and low-frequency autosomal SNPs. Based on the previous literature, I deem it unlikely that there would be a very large number of rare variants responsible for a substantial share of the heritability of the main phenotypes, but based on the studies in this thesis, we simply do not know. Also, it could be that allosomal variation (i.e., on the sex chromosomes) may explain part of the phenotypic variation, such as the observed difference in general risk tolerance between men and women. But we did not investigate variation on the sex chromosomes. Overall, the studies reported in this thesis are particularly focused on genetic variation rather than environment factors, while the heritability estimates strongly suggest that much of the variation in the main phenotypes is attributable to the environment. Thus, the lack of large-scale investigation of environmental factors could be considered a limitation of this thesis. Finally, I acknowledge that this thesis has barely scratched the surface of the genetic iceberg, and there are likely many thousands of loci left to discover in relation to the main phenotypes.

References – Chapter 6

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** 1001–1006 (2014).
2. Ripke, S. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18,** 497–511 (2013).
3. de Moor, M. H. M. *et al.* Meta-analysis of Genome-wide Association Studies for Neuroticism, and the Polygenic Association With Major Depressive Disorder. *JAMA Psychiatry* **72,** 642 (2015).
4. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Publ. Gr.* **47,** 284–290 (2015).
5. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50,** 906–908 (2018).
6. Collins, P. Y. *et al.* Grand challenges in global mental health. *Nature* **475,** 7–10 (2011).
7. Hyman, S. Mental health: Depression needs large human-genetics studies. *Nature* **515,** 189–191 (2014).
8. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15,** 765–776 (2014).
9. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11,** 446–450 (2010).
10. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
11. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
12. Wray, N. R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res. Hum. Genet.* **8,** 87–94 (2005).
13. Auer, P. L. & Lettre, G. Rare variant association studies: Considerations, challenges and opportunities. *Genome Med.* **7,** 1–11 (2015).
14. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.* **14,** 125–138 (2013).
15. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
16. Iulio, J. di *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* 82362 (2018). doi:10.1101/082362
17. Kitts, A., Phan, L., Ward, M. & Holmes, J. B. The Database of Short Genetic Variation (dbSNP). in *The NCBI Handbook [Internet]* (National Center for Biotechnology Information, 2013).
18. Alberts, B. *Molecular biology of the cell*. (Garland Science, Taylor and Francis Group, 2015).
19. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate

causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 1 (2018). doi:10.1038/s41576-018-0016-z

20. Yasunami, M., Kikuchi, I., Sarapata, D. & Yoshida, A. The human class I alcohol dehydrogenase gene cluster: Three genes are tandemly organized in an 80-kb-long segment of the genome. *Genomics* **7,** 152–158 (1990).
21. Stewart, M. J. *et al.* Promoters for the human alcohol dehydrogenase genes ADHI , ADH2 , and ADH3 : interaction of CCAAT / enhat , cer-binding protein with elements flanking the ADH2 TATA box The human ADHI , ADH2 , and ADH3 genes are closely related members of a gene family wh. **90,** 271–279 (1990).
22. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
23. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
24. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).
25. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0147-3
26. Rietveld, C. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–71 (2013).
27. Otowa, T. *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry* **21,** 1391 (2016).
28. Stetler, C. & Miller, G. E. Depression and hypothalamic-pituitary-adrenal activation: A quantitative summary of four decades of research. *Psychosom. Med.* **73,** 114–126 (2011).
29. Cueva, C. *et al.* Cortisol and testosterone increase financial risk taking and may destabilize markets. *Sci. Rep.* **5,** 1–16 (2015).
30. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9,** 436–447 (2016).
31. Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* **98,** 1–17 (2016).
32. Davies, N. M., Holmes, M. V & Smith, G. D. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362,** k601 (2018).
33. Heijmans, B. T. & Mill, J. Commentary: The seven plagues of epigenetic epidemiology. *Int. J. Epidemiol.* **41,** 74–78 (2012).
34. Cunliffe, V. T. The epigenetic impacts of social stress: how does social adversity become biologically embedded? *Epigenomics* **8,** 1653–1669 (2016).
35. Szyf, M., McGowan, P. & Meany, M. J. The Social Environment and the Epigenome. *Environ. Mol. Mutagen.* **49,** 46–60 (2008).
36. Kader, F., Ghai, M. & Maharaj, L. The effects of DNA methylation on human psychology. *Behav. Brain Res.* **346,** 47–65 (2018).

37. Flanagan, J. M. Chapter 3 – Epigenome-Wide Association Studies (EWAS): Past, Present, and Future. in *Cancer epigenetics - Risk Assessment, Diagnosis, Treatment, and Prognosis* (ed. Verma, M.) 51–64 (Springer, 2015).

38. Chen, D., Meng, L., Pei, F., Zheng, Y. & Leng, J. A review of DNA methylation in depression. *J. Clin. Neurosci.* **43,** 39–46 (2017).

39. Szyf, M. The early life social environment and DNA methylation - DNA methylation mediating the long-term impact of social environments early in life. 971–978 (2011). doi:10.461/epi.6.8.16793

40. Roth, T. L. Epigenetic mechanisms in the development of behavior: Advances, challenges, and future promises of a new field. *Dev. Psychopathol.* **25,** 1279–1291 (2013).

41. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).

42. Stringhini, S. *et al.* Life-course socioeconomic status and DNA methylation of genes regulating inflammation. *Int. J. Epidemiol.* **44,** 1320–1330 (2015).

43. Ehrenreich, B. Why are the poor blamed and shamed for their deaths? *The Guardian* 1–8 (2018).

44. Lundberg, J., Kristenson, M. & Starrin, B. Status incongruence revisited: Associations with shame and mental wellbeing. *Sociol. Heal. Illn.* **31,** 478–493 (2009).

45. World Health Organization. Mental Health Action Plan 2013-2020. *WHO Libr. Cat. DataLibrary Cat. Data* 1–44 (2013).

46. Daar, A. S. *et al.* Grand challenges in chronic non-communicable diseases. *Nature* **450,** 20–22 (2007).

47. Keller, M. C. Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biol. Psychiatry* **75,** 18–24 (2014).

48. Tielbeek, J. J. *et al.* Genome-Wide Association Studies of a Broad Spectrum of Antisocial Behavior. *JAMA Psychiatry* (2017).

49. Bustamante, C. D., Burchard, E. G. & De La Vega, F. M. Genomics for the world. *Nature* **475,** 163–165 (2011).

50. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).